GT-LOD3: LOD3 Semantic 3D Building Reconstruction Benchmark Dataset

Han Sae Kim[1], Olaf Wysocki[2], Ludwig Hoegner[3], Ksenia Bittner[4], Joshua Carpenter[5], Friedrich Fraundorfer[4,6], Arpan Kusari[7], Max Mehltretter[8], Franz Rottensteiner[8], Anna Schadl[9], Martin Weinmann[10], Jinha Jung[1,*]

[1] Lyles School of Civil and Construction Engineering, Purdue University, West Lafayette, IN, USA
[2] CV4DT, Department of Engineering, University of Cambridge, Cambridge, United Kingdom
[3] Faculty of Civil Engineering, Hochschule München University of Applied Sciences, Munich, Germany
[4] Remote Sensing Technology Institute, German Aerospace Center (DLR), Oberpfaffenhofen, Germany
[5] Department of Civil Engineering, The University of Akron, Akron, OH, USA
[6] Institute of Visual Computing, Graz University of Technology, Graz, Austria
[7] University of Michigan Transportation Research Institute, University of Michigan, Ann Arbor, MI, USA
[8] Institute of Photogrammetry and GeoInformation, Leibniz University Hannover, Hannover, Germany
[9] Faculty of Geoinformatics, Hochschule München University of Applied Sciences, Munich, Germany
[10] Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Keywords: Semantic reconstruction, Semantic segmentation, Low poly models, CAD benchmark dataset, CAD reconstruction, Urban digital twin

Abstract

Reconstructing semantic 3D building models at level of detail 3 (LOD3) is a long-standing challenge in photogrammetry, remote sensing, and computer vision. In contrast to conventional mesh-based representations, LOD3 models require watertight geometries along with semantic object-level facade elements. In GT-LOD3, we introduce the first LOD3 building benchmark dataset comprising point clouds and images paired with 32 ground truth (GT) LOD3 instances in two different countries. We also analyze the performance of selected baselines, complemented by a discussion on unresolved challenges. We are convinced that GT-LOD3 will facilitate the development of novel LOD3 reconstruction methods, enabling the widespread adoption of LOD3 models and, consequently, various downstream applications, ranging from energy demand estimation to automated driving function testing. We release the dataset as open-source and it can be accessed at https://github.com/gdslab/GT-LOD3-Benchmark

1. Introduction

Three-dimensional (3D) building models have become an essential foundation for modern digital-twin systems, supporting a wide range of applications such as energy performance studies, disaster risk assessment, and urban sustainability planning (Kolbe and Donaubauer, 2021). However, current engineering practice shows that the automatic 3D reconstruction of such models can be only conducted up to the level of detail (LOD) 2 that is characterized by simplified facades and complex roof structure (Haala and Kada, 2010, Lei et al., 2022, Wysocki et al., 2024a).

The next level of building detailing, level of detail 3 (LOD3), provides a richer representation by including facade elements such as windows and doors (Kolbe et al., 2021). These details are crucial for high-fidelity urban simulations, as they aid accurate prediction of heat transfer through facades, estimation of photovoltaic potential on building surfaces, assessment of flood and wind impacts, and even testing of autonomous driving functions and robot positioning. Yet, owing to reconstruction complexity and the scarcity of data acquired at street-level, LOD3 building models remain scarce, totaling only around 1,000 buildings worldwide, compared with approximately 216 million for LOD1 and LOD2 (Wysocki et al., 2024b, Wysocki et al., 2024a).

Recent automatic LOD3 reconstruction methods show great potential in tackling the involved challenges (Wysocki et al., 2023, Pantoja-Rosero et al., 2022, Huang et al., 2020, Hanke et al., 2025, Tang et al., 2025). However, to date, no LOD3 benchmark dataset with defined evaluation routines and paired sensor-to-model data is available. In effect, the introduced methods are only tested on few-instance datasets, limiting the evaluation to a selection of specific building types and sensor setups. Also, the comparison of methods’ efficiency is limited, as there is neither common data nor a common benchmark evaluation metric. To address these limitations, we present the following contributions:

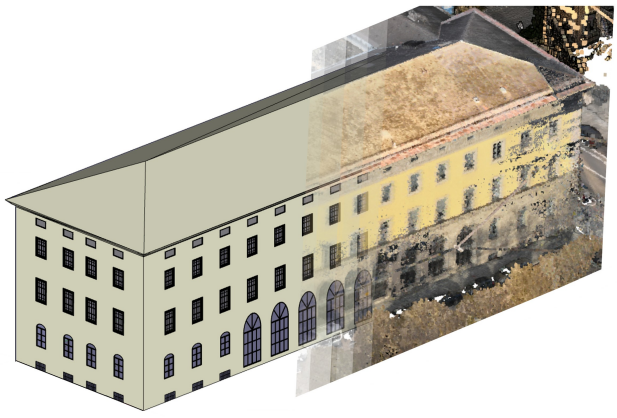


Figure 1. GT-LOD3 introduces a novel benchmark dataset of paired point clouds and GT LOD3 models for semantic 3D building reconstruction at LOD3.

- We introduce the first publicly available benchmark for

* Corresponding author: jinha@purdue.edu

LOD3 building reconstruction, GT-LOD3, that bridges facade-level segmentation and LOD3 modeling.

- The GT-LOD3 dataset integrates terrestrial laser scanning, mobile mapping, and oblique unmanned aerial system (UAS) photogrammetry across diverse architectural and environmental settings, all georeferenced and semantically aligned with LOD3 ground truth.
- A unified benchmarking protocol jointly assesses geometric accuracy and semantic completeness, enabling consistent and reproducible evaluation of LOD3 reconstruction methods.

Together, these contributions lay a foundation for the next generation of generalizable and reproducible LOD3 modeling research, offering the research community a unified and standardized platform for evaluating and advancing urban digital-twin reconstruction at the building scale. Figure 1 illustrates the proposed GT-LOD3 benchmark concept, which integrates high-resolution UAS point clouds with manually reconstructed LOD3 building models, enabling facade-level evaluation.

2. Related Work

2.1 Urban Benchmarks

Established open urban data benchmarks, such as Semantic3D.net (Hackel et al., 2017) and Toronto-3D (Tan et al., 2020), primarily focus on urban semantic segmentation of point clouds and do not include paired LOD3 models. Also, datasets targeting facade segmentation, such as TUM-FAÇADE (Wysocki et al., 2022c) or ArCH (Matrone et al., 2020), provide valuable facade annotations but focus narrowly on facade segmentation and lack paired LOD3 models. More recent initiatives, such as ZAHA (Wysocki et al., 2025b), introduce the hierarchical level of facade generalization (LoFG) for facade semantics, offering large-scale annotated mobile laser scanning (MLS) point clouds. In contrast, large multimodal benchmarks such as TUM2TWIN (Wysocki et al., 2025a) integrate terrestrial, aerial, and satellite data for comprehensive digital-twin analysis but cover only a relatively small area dominated by country-specific architectural styles. As a result, the current research landscape still lacks a standardized and diverse benchmark that enables reproducible evaluation and robust generalization across sensing modalities and architectural contexts. Table 1 provides an overview and comparison of publicly available datasets with potential relevance for LOD3 reconstruction research, highlighting the gap that motivates the development of GT-LOD3.

2.2 LOD3 Reconstruction Methods

Reconstructing semantic 3D buildings at LOD3 remains a major challenge in photogrammetry and computer vision. Traditionally, the creation of high-resolution LOD3 elements has relied on manual modeling (Chaidas et al., 2021). In recent years, research has increasingly focused on automating this process, leading to steady progress toward robust LOD3 reconstruction.

Despite these advances, LOD3 models remain rare in real-world applications. A key limitation is the lack of robustness of current methods when scaled to complex urban environments owing to modeling complexity and high data-demands for accurate reconstruction (Wysocki et al., 2023, Pantoja-Rosero et al., 2022, Pang and Biljecki, 2022, Hensel et al., 2019, Wang et al., 2024, Salehitangrizi et al., 2024, Hanke et al., 2025, Tang et al., 2025). Many approaches assume controlled acquisition conditions, requiring accurate co-registration of multimodal data and full, unobstructed coverage of each building. This often confines them to isolated building structures, such as standalone houses or blocks with unobstructed view captured via 360° UAS flights (Pantoja-Rosero et al., 2022, Huang et al., 2020, Kim et al., 2025) or oblique imagery (Xia et al., 2025), restricting broader applicability.

Other works have introduced the concept of conflict maps (CMs) for LOD3 reconstruction (Wysocki et al., 2022b). CMs are surface textures generated through scanner laser ray analysis with a prior LOD1 or LOD2 building model. This concept has evolved toward uncertainty-aware CMs that integrate Bayesian reasoning to fuse information from multiple modalities, such as segmented point clouds (Wysocki et al., 2022a) and segmented images (Wysocki et al., 2023). Further developments include inpainting-based techniques to address CM incompleteness (Froech et al., 2025) and using synthetic data to train robust CMs classifiers (Hanke et al., 2025). However, current CMs performance relies on the assumption of laser scanner, its position, and availability of low-LOD models to achieve semantic LOD3 reconstructions.

3. GT-LOD3: LOD3 Benchmark Dataset

This benchmark dataset consists of two subsets of buildings with three data types each: First, we provide photogrammetric UAS-based point clouds and images for Gold Coast and the Technical University of Munich (TUM) areas. Second, we provide ground-truth labels for point clouds to evaluate instance segmentation methods. And third, we provide ground-truth LOD3 building models of the test sites, containing 3D geometry of facades, facade elements such as windows, and roof structures. A visual overview of the GT-LOD3 dataset is shown in Figure 2.

3.1 Gold Coast dataset

3.1.1 Data acquisition The Gold Coast dataset was collected in the Gold Coast area of Lakewood, Ohio, USA, using a DJI Matrice 300 UAS equipped with a Zenmuse P1 RGB camera. The Zenmuse P1 employs a 45 MP full-frame CMOS sensor with a global mechanical shutter (up to 1/2000 s) and supports a 3 fps capture rate, providing distortion-free imagery during oblique flights. TimeSync 2.0 ensures centimeter-level alignment between sensor readings and GNSS/IMU observations, enabling high-accuracy photogrammetry (DJI Enterprise, 2023). Image acquisition was conducted using the “Smart Oblique Capture” mode, which allows automated collection of oblique imagery from multiple angles (nadir, front, back, left, and right) during a single flight. This mode reduces redundant image capture at the margins of the survey area, improves facade visibility, and enhances 3D reconstruction efficiency by up to 20–50% compared to traditional oblique systems (DJI Enterprise, 2021). Consecutive flight lines were designed to maintain a consistent 85% forward overlap and 85% sidelap, with a flight altitude of 120 meters above the tallest building, ensuring optimal conditions for photogrammetric reconstruction. A total of 4,705 oblique RGB images were acquired under this configuration, yielding a high-resolution dataset suitable for detailed facade interpretation and LOD3 building modeling. The collected oblique images were subsequently processed using Agisoft

Name	Country	Size	Aerial?	Terrestrial?	Footprints?	LOD3	LOD2	LOD1	Validation up to
TUM2TWIN	DE	city block	✓	✓	✓	✓	✓	✓	LOD3
Ingolstadt	DE	city block	✓	✓	✓	✓	✓	✓	LOD3
Poznań	PL	city	✓	∼	✓	∼	✓	✓	∼LOD3
Vienna	AT	city	✓	✓	✓	✗	✓	✓	LOD2
Prague	CZ	city	✓	✗	✓	✗	✓	✓	LOD2
Diekirch/Bastendorf	LU	city	✓	✗	✓	✗	✓	✓	LOD2
Riga	LV	city	✓	✗	✓	✗	✓	✓	LOD2
Espoo	FI	city	✓	✗	✓	✗	✓	✓	LOD2
Helsinki	FI	city	✓	✗	✓	✗	✓	✓	LOD2
Lyon	FR	city	✓	✗	✓	✗	✓	✓	LOD2
Berlin	DE	city	✓	✗	✓	✗	✓	✓	LOD2
Bremen/Bremerhaven	DE	city	✓	✗	✓	✗	✓	✓	LOD2
Potsdam	DE	city	✓	✗	✓	✗	✓	✓	LOD2
Rotterdam	NL	city	✓	✗	✓	✗	✓	✓	LOD2
Zurich	CH	city	✓	✗	✓	✗	✓	✓	LOD2
New York	US	city	✓	✗	✓	✗	✓	✓	LOD2
Bavaria	DE	region	✓	✗	✓	✗	✓	✓	LOD2
Brandenburg	DE	region	✓	✗	✓	✗	✓	✓	LOD2
North Rhine-Westphalia	DE	region	✓	✗	✓	✗	✓	✓	LOD2
Saxony	DE	region	✓	✗	✓	✗	✓	✓	LOD2
Saxony-Anhalt	DE	region	✓	✗	✓	✗	✓	✓	LOD2
Schleswig-Holstein	DE	region	✓	✗	✓	✗	✓	✓	LOD2
Thuringia	DE	region	✓	✗	✓	✗	✓	✓	LOD2
Netherlands	NL	country	✓	✗	✓	✗	✓	✓	LOD2
Poland	PL	country	✓	✗	✓	✗	∼	✓	∼LOD2
Estonia	EE	country	✓	✗	✓	✗	✓	✓	LOD2
Switzerland	CH	country	✓	✗	✓	✗	✓	✓	LOD2
USA	US	country	✓	✗	✓	✗	✗	✓	LOD1
$\sum$ 27									

Table 1. List of datasets revealing potential for training and validation of building reconstruction methods, where the feature is present (✓), absent (✗), and partially available (∼). Table based on (Wysocki et al., 2024a).

Metashape Professional, where an orthomosaic image, digital surface model (DSM), and 3D point cloud were generated as intermediate products for LOD3 reconstruction. All geospatial data products were projected in the EPSG:6549 coordinate reference system (NAD83 / Ohio South [ftUS]).

3.1.2 LOD3 Ground Truth Modeling From the surveyed region, five representative buildings were selected for LOD3 reconstruction. These buildings were chosen to cover a variety of architectural forms and boundary shapes, with clearly distinguishable windows and facade elements. The manual LOD3 reconstruction workflow follows the procedures described by (Kim et al., 2025, Wysocki et al., 2025a). Windows and doors were manually annotated from the UAS imagery and projected onto the corresponding LOD2 building models to identify precise corner points. Using SketchUp with the CityEditor plugin, the LOD3 models were reconstructed by referencing these control points and aligning them with the facade geometry. The resulting models accurately represent the facade structures and opening geometries, serving as ground-truth data for algorithm benchmarking.

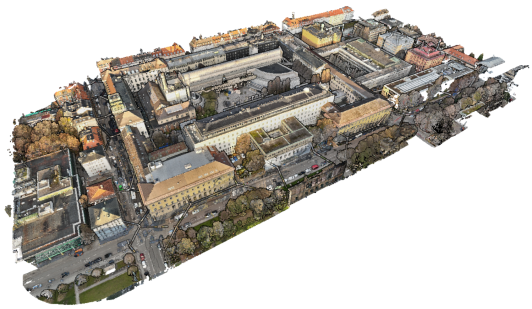
3.2 TUM UAS dataset

3.2.1 Data acquisition The downtown campus of the TUM has been recorded with a large variety of sensors in the last years including several MLS campaigns, thermal infrared and optical RGB images from the ground and from UAS (Wysocki et al., 2025a). ZAHA (Wysocki et al., 2025b) introduces the LoFG based on MLS datasets: A novel hierarchical classification system based on international urban modeling standards. This approach ensures compatibility with real-world architectural challenges while enabling a standardized comparison of segmentation methods. As part of this effort, a large 3D facade semantic segmentation dataset is presented, comprising 601 million annotated points across five classes in LoFG2 and 15 classes in LoFG3: wall, window, door, molding, deco, column, arch, stairs, ground surface, terrain, roof, blinds, interior, and other.

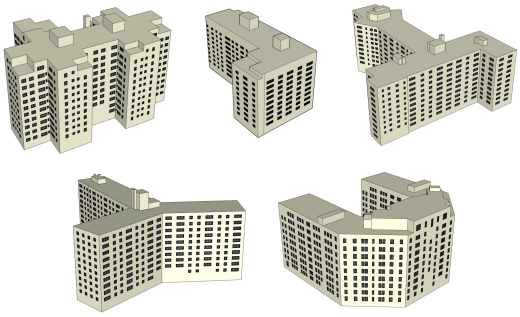
In this dataset, we concentrate on the UAS images to include also roof structures and inner yards. During the survey, 1104 images with a resolution of 5,280 × 3,956 pixels are taken by a Zenmuse L2 integrated RGM mapping camera (4/3 CMOS) mounted on a DJI Matrice 350 RTK drone. The data is collected in nadir mode using automatic flight missions at an altitude of 75 m above ground. Additional acquisitions are collected in manually operated flights in oblique view to capture the facades of the inner and outer campus area. The average ground sampling distance (GSD) of images is 1.6 cm, and these images are georeferenced through the real-time kinematic (RTK) global navigation satellite system (GNSS) measurements and inertial measurement unit (IMU) of the UAS. Images including nadir and oblique acquisition geometry are filtered by removing recognizable persons and license plates, finally totaling a repository of 962 images. These images can be used for coloring laser scans, orthophoto generation, and photogrammetric reconstruction (Anders et al., 2024).

3.2.2 LOD3 Ground Truth Modeling A LOD3 city geography markup language (CityGML) model of the TUM campus and surrounding buildings was generated manually as

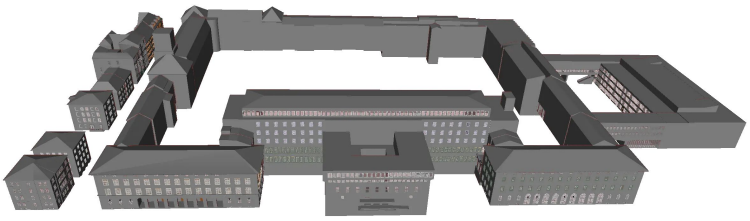
(a) Gold Coast area reconstructed from oblique UAS imagery.



(b) Area of the TUM city campus in Munich, Germany



(c) Manually reconstructed LOD3 models of five buildings of the Gold Coast area.



(d) Manually generated LOD3 model of the TUM city campus

Figure 2. An overview of the Gold Coast and TUM photogrammetric UAS-based point clouds and manually reconstructed LOD3 models.

ground truth, consisting of 28 building instances in total with the 15 classes already mentioned. The main difference to LOD1 and LOD2 models pertains to correcting the geometry in presence of overhangs (roof geometry) and in enriching facades in openings (i.e., around 500 windows and doors) and building installation (e.g., stairs) if exceeding the threshold of intrusion or extrusion by 10 cm. The 3D measurements of combined proprietary point clouds (3D Mapping Solutions, 2023) and TUM-MLS-16 (Zhu et al., 2020) were used for modeling. Additionally, a 3D library of computer-aided design (CAD) models of facade elements is created to allow 3D elements modeling. The city model contains not only the 3D building models, but also semantic models of the street space at lane-level granularity and of the vegetation. The 3D photogrammetric points are not labeled directly, but we transfer the labels of ZAHA using nearest neighbor. For parts of the photogrammetric point cloud not recorded in TUM-MLS-16, like inner yards and roofs, we derive the labels from the closest element in the LOD3 CityGML model.

4. Experiments

4.1 Baseline LOD3 reconstruction method

As illustrated in Figure 3, to provide a reproducible reference for evaluating the proposed benchmark, we develop a baseline LOD3 reconstruction framework that integrates point-cloud semantic segmentation, facade-level projection, and opening-to-geometry conversion. The pipeline consists of three main stages: (1) semantic labeling of photogrammetric point clouds, (2) facade-aligned 2D rasterization, and (3) projection-based inference of facade openings for LOD3 model construction.

4.2 Point-Cloud Semantic Segmentation

The photogrammetric point clouds derived from UAS imagery are first semantically segmented into five classes: ground, wall, roof, window, and other. For this task, we adopt the Point Transformer (Zhao et al., 2021) and Point Transformer V3 (Wu et al., 2024) architecture due to its strong generalization across diverse facade conditions.

- **Weak Supervision for the Gold Coast Data:** For the Gold Coast dataset, instead of manually annotating large photogrammetric point clouds, we generate weak labels by referencing manually reconstructed LOD3 building models and transferring their semantics onto the aligned UAS point clouds. Using this weakly supervised dataset, a Point Transformer model is trained from scratch such that the baseline adapts to the distinct architectural and environmental characteristics of the Gold Coast buildings.

- **ZAHA-trained Model for TUM2TWIN UAS dataset:** For the TUM buildings, segmentation is performed using a Point Transformer model pretrained on the ZAHA benchmark dataset, which contains approximately 601 million semantically annotated facade points. The dataset includes a wide range of facade types, such as residential buildings, educational facilities, cultural heritage facades, and underpass structures. The validation and test subsets reflect realistic facade variations, while the remaining facades serve as the training set. Additional architectural and annotation details are available in the original ZAHA publication (Wysocki et al., 2025b).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-293-2026 | © Author(s) 2026. CC BY 4.0 License.

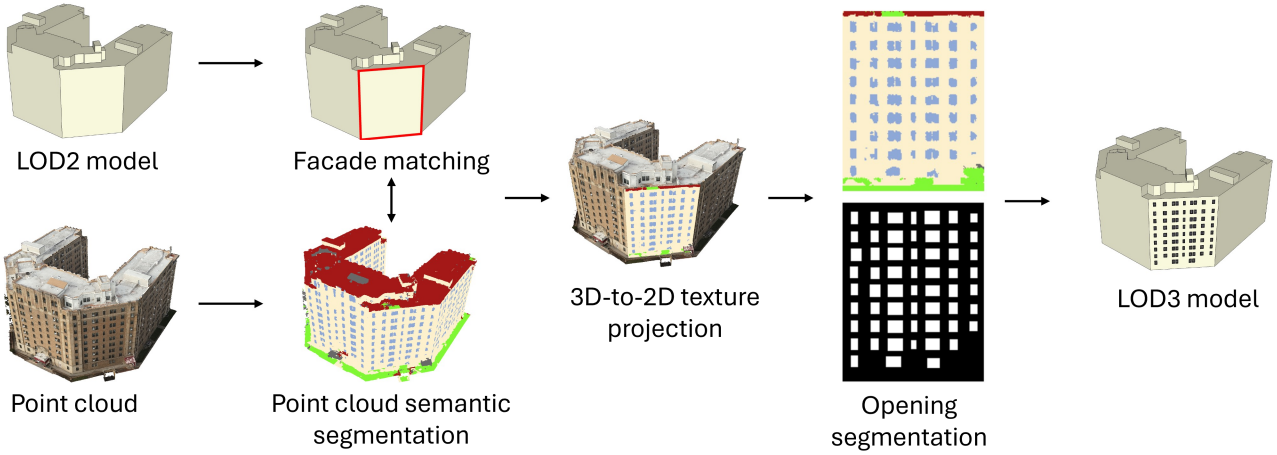


Figure 3. Workflow of the baseline LOD3 reconstruction method. The photogrammetric point cloud is semantically segmented into five classes: ground, wall, roof, window, and other. Using the LOD2 model, facades are clipped and projected onto a 2D plane for opening estimation, which is then reprojected onto the LOD2 geometry to generate the final LOD3 model.

- **Training Configuration:** All experiments are conducted on an NVIDIA GeForce RTX 4090 GPU with 24 GB VRAM. The model is trained for 100 epochs, using a batch size of 32 and each batch containing 8,192 points. We use stochastic gradient descent with a momentum of 0.1 and an initial learning rate of 0.001. To ensure spatial locality in training batches, we employ a Z-order (Morton) based batch design strategy, which is computationally efficient for organizing point clouds along a space-filling curve (Morton, 1966). This approach aligns with recent evidence that serialization-based neighborhood construction provides a substantially more efficient alternative to KNN, enabling faster training without degrading accuracy (Wu et al., 2024). Training, validation, and test subsets are constructed such that all five semantic classes appear across all splits. For testing, facade-level subsets are generated by clipping the point cloud using the corresponding facade geometry.

4.3 Facade-Aligned 2D Projection and Class Map Generation

Following semantic segmentation, the LOD2 model serves as the structural reference for isolating individual facades. For each building, the corresponding LOD2 facade polygons are extracted and used to clip the segmented point cloud into facade-specific subsets. Each facade-level point cloud is then orthogonally projected onto a local 2D plane. A rasterized class map is generated, where each pixel represents the majority semantic label of the points falling within its area. The rasterization resolution was selected to ensure that each pixel aggregates, on average, two to three projected points from the photogrammetric point cloud. This design choice stabilizes the majority-voting process used during class map generation, mitigates sensitivity to local sampling noise, and maintains consistency with the nominal coordinate precision defined in the LOD2 model.

4.4 Opening Extraction and LOD3 Model Construction

From each facade class map, the window class is converted into a binary mask. The mask is processed using morphological operators to regularize window shapes into rectangular forms, after which it is vectorized, transformed back into the facade’s 3D coordinate system, and projected onto the corresponding LOD2 geometry. By combining the LOD2 model with the reconstructed opening boundaries, the baseline method produces LOD3 models that include explicit window geometries for each facade.

4.5 Evaluation protocols

The evaluation procedure employed in this study follows the methodology introduced in Scan2LoD3 and CM2LoD3 (Wysocki et al., 2023, Hanke et al., 2025), which provide a structured protocol for quantifying the geometric and semantic quality of LOD3 building reconstruction. Consistent with that framework, our evaluation is performed at two complementary levels. At the geometric level, we assess the spatial fidelity of reconstructed facade openings by comparing the projected 2D opening masks against ground-truth masks derived from LOD3 models. At the instance level, we evaluate the reconstruction pipeline’s ability to correctly detect individual window objects.

- **Geometric Accuracy:** To assess geometric accuracy, each predicted opening mask is projected onto a facade-aligned local coordinate plane, where it is rasterized at a consistent resolution. Ground-truth masks are generated in the same coordinate frame by projecting the LOD3 window geometries defined in the CityGML file and rasterizing their polygon surfaces. Because both predicted and reference masks lie in the same facade plane, pixel-level comparisons can be made directly. Geometric fidelity is quantified using precision, recall, and F1-score. In addition, we compute the intersection over union (IoU) between predicted and ground-truth masks, which provides a holistic measure of the spatial overlap between the reconstructed opening region and the corresponding LOD3 reference geometry. These metrics are particularly meaningful for facade openings, which occupy only a small fraction of the facade area and are thus sensitive to localization errors.

- **Detection Rate:** To evaluate object-level reconstruction quality, we adopt an instance-based matching procedure. First, connected-component analysis is applied to both predicted and ground-truth masks, producing a set of individual window instances for each facade. For every pair of predicted and ground-truth objects, a pairwise IoU is

computed. Using the resulting IoU matrix, a cost matrix defined as 1-IoU is constructed, and the Hungarian algorithm is applied to determine the optimal one-to-one assignment between predicted and reference windows. A matched pair is considered a true positive if its IoU exceeds 0.5. Predicted instances without valid matches are counted as false positives, and unmatched ground-truth windows are counted as false negatives.

This combination of pixel-level and instance-level metrics allows for a comprehensive evaluation of LOD3 facade reconstruction. The pixel-level analysis quantifies the geometric accuracy of opening shapes and extents, while the instance-level evaluation reflects the correctness of object detection, including the number, placement, and distinctness of individual openings.

5. Results and Discussion

5.1 Semantic Segmentation Performance

The semantic segmentation stage provides the foundation for the subsequent LOD3 reconstruction, and its performance directly impacts the fidelity of the opening geometry. The qualitative results in Figure 4 demonstrates that across the six test facades, the Point Transformer generally preserves the dominant facade components, including wall surfaces, roof structures, and ground regions. In comparison with the ground truth, the predicted wall and roof labels show relatively coherent spatial distributions, indicating that the model effectively captures large and geometrically continuous classes. However, the window regions exhibit more fragmented and incomplete predictions, particularly in the edges of the window layouts. Some window points are misclassified as wall, resulting in partially missing openings, while a small number of roof and ground points are confused near facade boundaries and occluded areas. These errors are visually consistent with the lower recall observed for the window class in Table 2. Since windows are the primary elements used to define openings in the subsequent LOD3 reconstruction, such under-segmentation can directly affect the completeness and geometric accuracy of reconstructed facade openings. As shown in Table 2, the Point Transformer model achieves strong performance for the major structural classes—particularly wall and roof, with F1 scores of 89.4% and 87.9% for the Gold Coast dataset and 95.4% and 90.5% for the TUM UAS dataset respectively. Window classification remains more challenging. The window class achieves an F1-score for Gold Coast of 72.9%, with a noticeable drop in recall (69.7%) relative to precision (76.3%). Similar results are for the TUM UAS data set with an F1 score of 72.3% with a recall of 65.6% and a precision of 80.5%. This discrepancy suggests that the model tends to under-segment window regions, which can be attributed to two primary factors: (1) the significantly smaller spatial footprint of windows relative to other classes, leading to class imbalance; and (2) the weakly supervised nature of the training labels, which inherently introduces some noise in boundary regions. These effects are consistent with findings in existing facade segmentation benchmarks and reveal the need for more targeted augmentation or specialized modules to handle small-structure classes.

5.2 Geometric Accuracy of Opening Reconstruction

The geometric consistency of reconstructed openings, evaluated through 2D facade-aligned projection, is reported in Table 3.

Class	Points (K)	Precision (%)	Recall (%)	F1 (%)	IoU (%)
Ground	93.2	63.1	58.5	60.7	43.6
Wall	2246.7	87.7	91.1	89.4	80.8
Roof	485.3	88.3	87.3	87.8	78.3
Window	659.7	76.3	69.7	72.8	57.3
Other	78.7	62.2	49.3	55.0	37.9
Total	3563.5	75.5	71.2	73.2	59.6

Table 2. A summary of semantic segmentation result of the test facades point cloud.

The baseline method demonstrates varying performance across facades. For the most structurally simple and noise-free facades (Facades IV and V), the approach achieves IoU scores of 76.8% and 74.1%, respectively. These facades correspond to buildings with clean, high-contrast window features and minimal occlusion, favoring robust contour extraction. In contrast, Facades I and II present more challenging conditions, including irregular textures, shadows, and partial occlusions. These cases yield lower IoU scores (45.0% and 55.5%) and reduced F1-scores (62.1% and 71.4%). The drop in performance indicates that segmentation inconsistencies and incomplete point density along the facade surface propagate through the projection and shape-regularization steps, ultimately impacting geometric accuracy. Nevertheless, the overall performance across the five facades demonstrates that the baseline method provides a stable starting point for future LOD3 reconstruction research.

Metric	Precision (%)	Recall (%)	F1 (%)	IoU (%)
Facade I	69.3	56.2	62.1	45.0
Facade II	81.2	63.8	71.4	55.5
Facade III	73.7	71.9	72.8	57.2
Facade IV	93.4	81.2	86.9	76.8
Facade V	94.8	77.3	85.1	74.1
Facade VI	57.7	46.4	51.5	34.6
Total	78.4	66.13	71.6	57.2

Table 3. An evaluation of semantic segmentation of the resulting LOD3 model.

5.3 Window Instance Detection Performance

Table 4 reports the results of instance-level window detection using the Hungarian matching protocol. Consistent with the geometric evaluation, facades exhibiting clear architectural structure show superior performance. Facades IV and V achieve F1-scores of 98.7% and 76.2%, respectively, demonstrating that the pipeline reliably detects the correct number of openings and accurately localizes them in these structured environments. More complex facades, particularly Facades II and III, exhibit substantially lower detection rates. The large number of false positives and false negatives suggests that even minor geometric inconsistencies during segmentation or rasterization can fragment or merge window shapes, making instance-level matching difficult. Facade I, in particular, showed a notably low number of true positives while simultaneously producing many false positives and false negatives. Visual inspection of the overlaid ground-truth and predicted masks reveals a slight spatial misalignment, leading to low IoU values and inflating both FP and FN counts. This outcome highlights the sensitivity of instance-level metrics to small spatial deviations and emphasizes the need for reconstruction methods that better preserve window boundaries, especially in facades with low texture or uneven illumination. The facade-wise variability in instance-level performance is visualized in Figure 5, which presents the IoU and F1-score for each facade, highlighting how architectural complexity and facade composition influ-

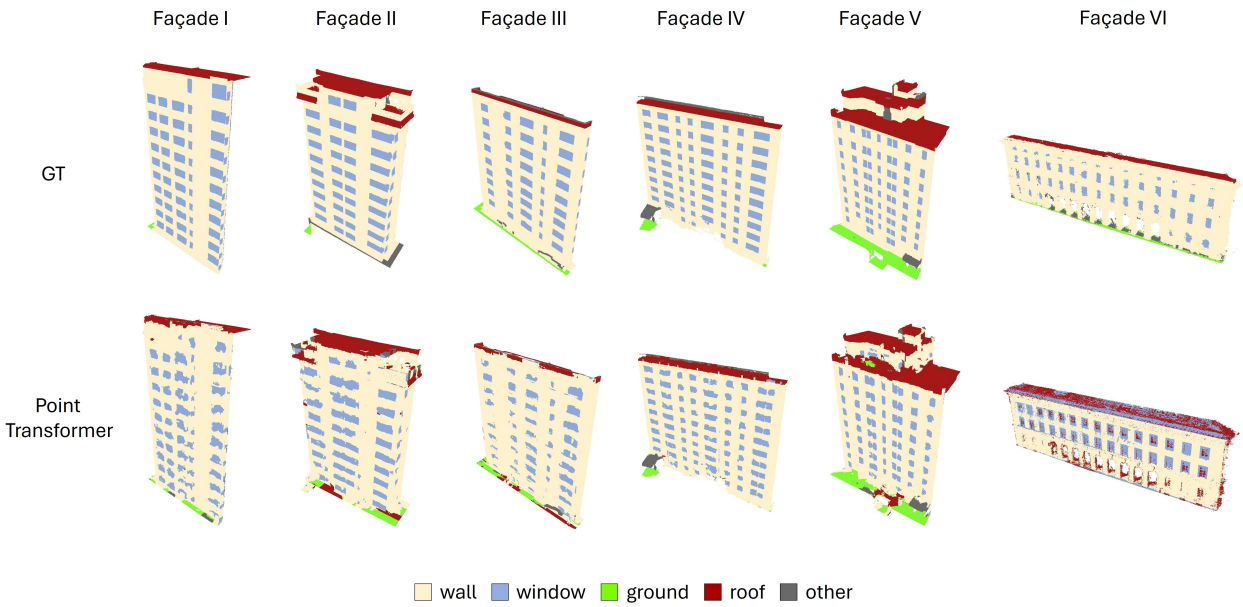


Figure 4. Qualitative comparison of semantic segmentation results on six test facades. The first row shows the ground truth labels, and the second row shows the predictions generated by the Point Transformer model.

Metric	Windows	TP	FP	FN	P (%)	R (%)	F1 (%)
Facade I	51	13	34	38	27.7	25.5	26.5
Facade II	38	16	15	22	51.6	42.1	46.4
Facade III	45	30	16	15	65.2	66.7	65.9
Facade IV	77	77	2	0	97.5	100.0	98.7
Facade V	68	48	10	20	82.8	70.6	76.2
Facade VI	57	11	19	46	36.7	19.3	25.3
Total	336	195	96	141	60.3	54.0	**56.5**

Table 4. Window detection statistics per facade with precision/recall/F1 in percent.

ence reconstruction accuracy.

5.4 Overall Observations and Implications

The results collectively show that the proposed dataset and evaluation protocol effectively reveal the strengths and limitations of the baseline LOD3 reconstruction approach. The baseline method performs well in controlled, structurally clean facade settings but degrades in more complex conditions where point density, illumination, and facade texture introduce challenges. Importantly, the performance gaps observed across facades underscore the value of the dataset’s multi-environment design; the benchmark reveals generalization weaknesses that do not emerge in evaluations limited to a single site or a homogeneous architectural style. The trends across semantic segmentation, geometric mask accuracy, and instance-level window detection consistently indicate that improving the robustness of opening segmentation is essential for advancing LOD3 reconstruction. The benchmark dataset introduced in this work provides a foundation for developing these improvements by offering paired sensor-to-model data and consistent benchmarking across diverse architectural contexts. Although the GT-LOD3 dataset currently comprises 32 building instances, it is essential to emphasize that it is primarily designed as a high-fidelity evaluation benchmark rather than a large-scale training corpus for deep learning. Given that LOD3 models are exceptionally scarce, our dataset provides a critical, rigorous testing ground for LOD3 reconstruction methods specifically utilizing UAS-based photogrammetric point clouds.

6. Conclusion

This study presents GT-LOD3, a UAS-derived point cloud dataset paired with manually reconstructed LOD3 models, establishing a new benchmark for facade-level building reconstruction. Unlike existing datasets that focus primarily on facade semantic segmentation or LOD2 geometry reconstruction, the proposed benchmark provides both semantic and geometric evaluation for high-detail and semantic-rich LOD3 reconstruction. By integrating photogrammetric UAS point clouds with manually constructed LOD3 building models, the dataset enables rigorous assessment of opening-level reconstruction performance under diverse architectural styles and environmental conditions.

A baseline reconstruction pipeline was introduced to demonstrate the utility of the benchmark dataset. The evaluation results highlight both the potential and the challenges of current point-cloud–based reconstruction methods. Facades with regular, well-defined window structures are reconstructed with high geometric consistency, whereas complex facades, occluded regions, or those exhibiting subtle geometric variations exhibit significant performance degradation. Importantly, the facade-wise variation in detection and geometric accuracy underscores the benchmark’s value: it exposes generalization gaps that would remain hidden in datasets limited to a single region or architectural style.

Overall, this work contributes (i) a UAS point cloud paired with detailed LOD3 ground truth, (ii) a standardized evaluation protocol for facade semantic reconstruction, and (iii) an initial baseline that can serve as a reference for future research. We envision that GT-LOD3 will support the development of more robust, generalizable, and semantically complete LOD3 reconstruction methods, thereby advancing the state of the art in urban digital-twin modeling.

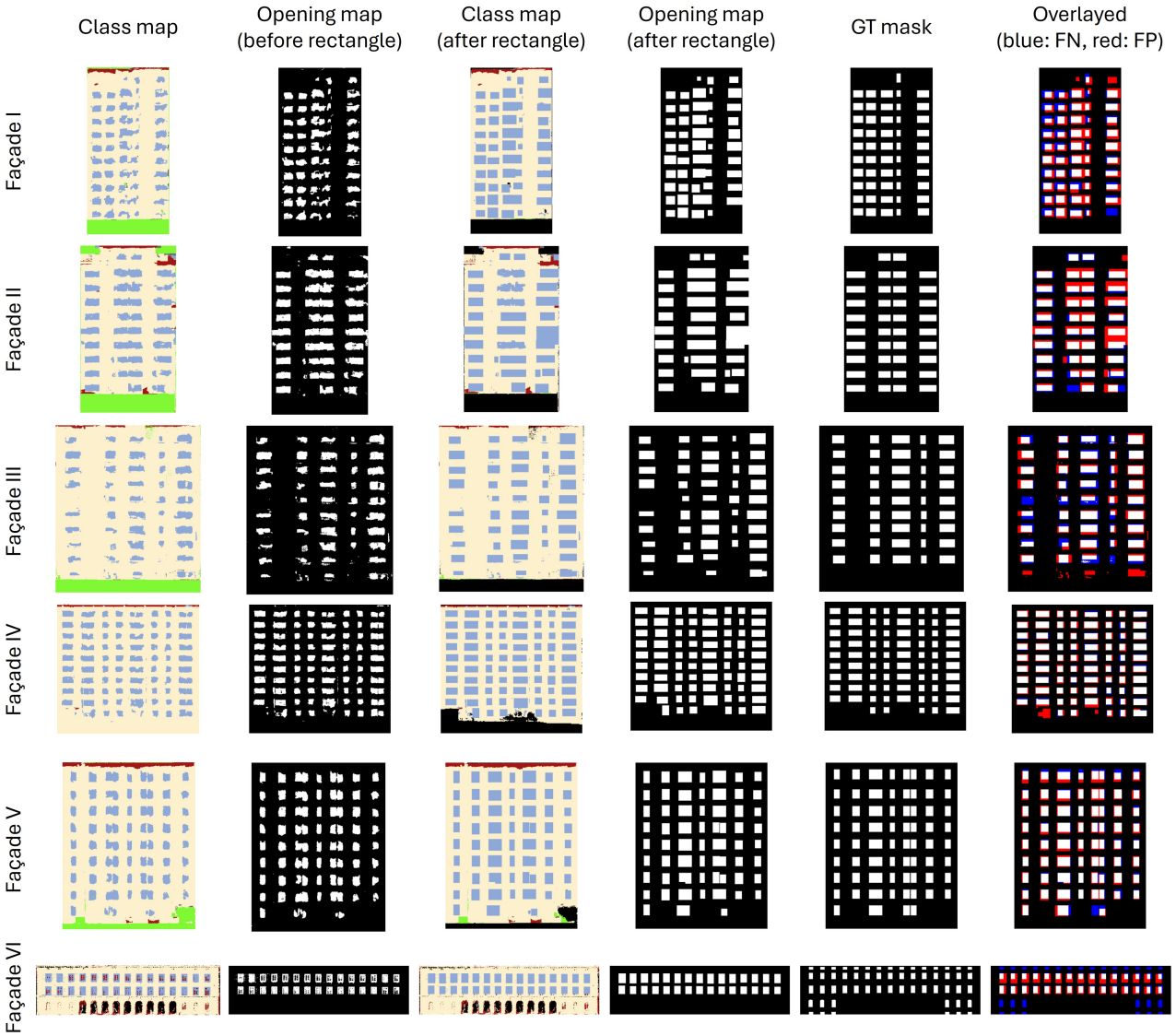


Figure 5. Evaluation over six test facades of the benchmark. 2D facade-aligned geometric accuracy and instance-level detection rate were calculated using 2D projected opening masks and GT masks from LOD3 model.

7. Data Availability

To facilitate further research in semantic 3D building reconstruction, the GT-LOD3 dataset—including the UAS-derived point clouds, images, and manually reconstructed LOD3 ground truth models—is made publicly available. The dataset can be accessed at . The data is released under the Creative Commons 4.0 International (CC BY 4.0) license. The data can be accessed https://github.com/gdslab/GT-LOD3-Benchmark,

References

3D Mapping Solutions, 2023. MoSES mobile mapping platform - technical details. `https://www.3d-mapping.de/ueber-uns/unternehmensbereiche/data-acquisition/unser-vermessungssystem/`. Accessed: 2023-01-30.

Anders, K., Wang, J., Chang, M., Letard, M., Schulte, F., Winiwarter, L., 2024. Terrestrial and UAV laser scanning point clouds of TUM Campus Ottobrunn.

Chaidas, K., Tataris, G., Soulakellis, N., 2021. Seismic damage semantics on post-earthquake LOD3 building models generated by UAS. *ISPRS International Journal of Geo-Information*, 10(5). https://www.mdpi.com/2220-9964/10/5/345.

DJI Enterprise, 2021. The zenmuse p1’s smart oblique capture is revolutionizing oblique aerial photography and 3d mapping. `https://enterprise-insights.dji.com/blog/smart-oblique-capture`. Accessed: 2025-11-15.

DJI Enterprise, 2023. Zenmuse p1 – full-frame 45 mp photogrammetry camera product specifications. `https://enterprise.dji.com/zenmuse-p1/specs`. Accessed: 2025-11-15.

Froech, T., Wysocki, O., Xia, Y., Xie, J., Schwab, B., Cremers, D., Kolbe, T. H., 2025. FacaDiffy: Inpainting Unseen Facade Parts Using Diffusion Models. *arXiv preprint arXiv:2502.14940*.

Haala, N., Kada, M., 2010. An update on automatic 3D building reconstruction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6), 570 - 580.

Hackel, T., Savinov, N., Ladicky, L., Wegner, J. D., Schindler, K., Pollefeys, M., 2017. Semantic3D.net: A new Large-scale Point Cloud Classification Benchmark. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-1/W1, 91–98.

Hanke, F., Bieringer, A., Wysocki, O., Jutzi, B., 2025. CM2LoD3: Reconstructing LoD3 Building Models Using Semantic Conflict Maps. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 81–88.

Hensel, S., Goebbels, S., Kada, M., 2019. Facade reconstruction for textured LOD2 CityGML models based on deep learning and mixed integer linear programming. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W5, 37–44.

Huang, H., Michelini, M., Schmitz, M., Roth, L., Mayer, H., 2020. LOD3 building reconstruction from multi-source images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2020, 427–434.

Kim, H. S., Carpenter, J., Jung, J., 2025. Automated lod-3 reconstruction using oblique uav images. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, Copernicus Publications Göttingen, Germany, 797–804.

Kolbe, T. H., Donaubauer, A., 2021. Semantic 3D city modeling and BIM. W. Shi, M. F. Goodchild, M. Batty, M.-P. Kwan, A. Zhang (eds), *Urban Informatics*, Springer Singapore, Singapore, 609–636.

Kolbe, T. H., Kutzner, T., Smyth, C. S., Nagel, C., Roensdorf, C., Heazel, C., 2021. OGC City Geography Markup Language (CityGML) Part 1: Conceptual Model Standard v3.0.

Lei, B., Stouffs, R., Biljecki, F., 2022. Assessing and benchmarking 3D city models. *International Journal of Geographical Information Science*, 0(0), 1-22.

Matrone, F., Lingua, A., Pierdicca, R., Malinverni, E. S., Paolanti, M., Grilli, E., Remondino, F., Murtiyoso, A., Landes, T., 2020. A benchmark for large-scale heritage point cloud semantic segmentation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2020, 1419–1426.

Morton, G. M., 1966. A computer oriented geodetic data base and a new technique in file sequencing.(1966).

Pang, H. E., Biljecki, F., 2022. 3D building reconstruction from single street view images using deep learning. *International Journal of Applied Earth Observation and Geoinformation*, 112, 102859.

Pantoja-Rosero, B. G., Achanta, R., Kozinski, M., Fua, P., Perez-Cruz, F., Beyer, K., 2022. Generating LoD3 building models from structure-from-motion and semantic segmentation. *Automation in Construction*, 141, 104430.

Salehitangrizi, A., Jabari, S., Sheng, M., Zhang, Y., 2024. 3D Modeling of Façade Elements Using Multi-View Images from Mobile Scanning Systems. *Canadian Journal of Remote Sensing*, 50(1), 2309895.

Tan, W., Qin, N., Ma, L., Li, Y., Du, J., Cai, G., Yang, K., Li, J., 2020. Toronto-3d: A large-scale mobile lidar dataset for semantic segmentation of urban roadways. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 797–806.

Tang, W., Li, W., Liang, X., Wysocki, O., Biljecki, F., Holst, C., Jutzi, B., 2025. Texture2lod3: Enabling lod3 building reconstruction with panoramic images. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2016–2026.

Wang, Y., Jiao, W., Fan, H., Zhou, G., 2024. A framework for fully automated reconstruction of semantic building model at urban-scale using textured LoD2 data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 216, 90–108.

Wu, X., Jiang, L., Wang, P.-S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H., 2024. Point transformer v3: Simpler faster stronger. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4840–4851.

Wysocki, O., Grilli, E., Hoegner, L., Stilla, U., 2022a. Combining visibility analysis and deep learning for refinement of semantic 3D building models by conflict classification. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-4/W2-2022, 289–296.

Wysocki, O., Hoegner, L., Stilla, U., 2022b. Refinement of semantic 3D building models by reconstructing underpasses from MLS point clouds. *International Journal of Applied Earth Observation and Geoinformation*, 111, 102841.

Wysocki, O., Hoegner, L., Stilla, U., 2022c. TUM-FACADE: Reviewing and enriching point cloud benchmarks for facade segmentation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVI-2/W1-2022, 529–536. https://isprs-archives.copernicus.org/articles/XLVI-2-W1-2022/529/2022/.

Wysocki, O., Schwab, B., Beil, C., Holst, C., Kolbe, T. H., 2024a. Reviewing Open Data Semantic 3D City Models to Develop Novel 3D Reconstruction Methods. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 493–500.

Wysocki, O., Schwab, B., Biswanath, M. K., Greza, M., Zhang, Q., Zhu, J., Froech, T., Heeramaglore, M., Hijazi, I., Kanna, K. et al., 2025a. TUM2TWIN: Introducing the Large-Scale Multimodal Urban Digital Twin Benchmark Dataset. *arXiv preprint arXiv:2505.07396*.

Wysocki, O., Schwab, B., Willenborg, B., Knezevic, M., 2024b. Awesome CityGML. `https://github.com/OloOcki/awesome-citygml`. Accessed: 2024-01-30.

Wysocki, O., Tan, Y., Froech, T., Xia, Y., Wysocki, M., Hoegner, L., Cremers, D., Holst, C., 2025b. Zaha: Introducing the level of facade generalization and the large-scale point cloud facade semantic segmentation benchmark dataset. *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 7648–7658.

Wysocki, O., Xia, Y., Wysocki, M., Grilli, E., Hoegner, L., Cremers, D., Stilla, U., 2023. Scan2LoD3: Reconstructing semantic 3D building models at LoD3 using ray casting and Bayesian networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 6547–6557.

Xia, Y., Gao, W., Stoter, J., 2025. Enriching LoD2 Building Models with Façade Openings Using Oblique Imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 225–232.

Zhao, H., Jiang, L., Jia, J., Torr, P. H., Koltun, V., 2021. Point transformer. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 16259–16268.

Zhu, J., Gehrung, J., Huang, R., Borgmann, B., Sun, Z., Hoegner, L., Hebel, M., Xu, Y., Stilla, U., 2020. TUM-MLS-2016: An annotated mobile LiDAR dataset of the TUM City Campus for semantic point cloud interpretation in urban areas. *Remote Sensing*, 12(11), 1875.